The Dangers of Parametrics

Or how we use models to fool ourselves and mislead our customers

Andy Prince

Introduction

From its roots in the World War II application of learning curve theory to the problem of aircraft production, parametric cost modeling and parametric cost estimating has grown to be a valuable tool for predicting the cost of complex and unique systems. Commercial parametric cost models are available to estimate almost any system, large industrial organizations and government agencies invest significant resources in developing and maintaining parametric estimating tools and capabilities, and professional organizations such as ICEAA provide a forum for exchanging ideas and promoting the growth of the profession as well as estimating professionals.

Over the decades parametric cost models have grown in capability and sophistication. RCA PRICE marketed the first general purpose parametric cost model in 1975. PRICE proved to be so popular that it spawned the PRICE User's Group which led directly to the creation of the International Society of Parametric Analysts (ISPA), a forerunner of ICEAA. Other commercial tools followed, including SEER by Galorath Incorporated and the now defunct FAST model. Models for estimating software cost (COCOMO) or complex systems (COSYSMO) were developed by universities and became popular tools. Government cost models, such as the NASA Air Force Cost Model (NAFCOM), were developed and distributed for free, enabling cost estimators around the world to parametrically estimate the cost of space flight hardware systems.

So, let's be honest with ourselves: we love our cost models. Most parametric cost estimators become masters of their models, able to explain (and justify) how specific model settings change a cost estimate and why. We can make our models produce whatever number we need, often supported by technical professionals who are usually very willing to help us make sure the model settings are consistent with their design assumptions. The end result is an estimate that is logically consistent with the technical and programmatic characteristics of the system. Not everyone may like our answer, but used properly our models give us answers that are credible, supportable, and defendable.

But what if our models are not solving our estimating problems, but instead are the source of our problems? What if our trust in our models is misplaced? What if our models are not as good as we think they are? What if we are substituting being good model users for being good cost estimators?

The remainder of this paper addresses these questions. We will look at what makes a good cost model and what makes a poor one. We will examine how we use subjective parameters to give us outstanding model behavior, and how those same subjective parameters can hurt our ability to provide accurate estimates. We will take time to understand how human psychology biases us in our quest to develop good cost models. We will take a look at a simple cost model, and a simple cost estimate to see what can go wrong when we trust the model over trusting the data. Finally, we will identify specific actions we can take to improve our models and how we use them.

The purpose of this paper is not to denigrate cost models. *Au contraire!* Cost models are vital to how we do our job. We need powerful, sophisticated parametric cost models. Rather, the purpose of this paper is to draw a distinction between cost estimating and cost analysis. To demonstrate that by becoming

better cost analysts, we can more effectively use the power we have available to us in our cost models to be better cost estimators.

Building Parametric Cost Models

Those of us who have been in the business for a few years know how to build a parametric cost model. You gather some data, either data from your company or government organization, or from somebody else's company or government organization. You normalize the data to make it as homogeneous as possible, then you start looking for relationships between and cost and various technical or programmatic parameters. Once we have found some likely candidates, we start doing regression analysis or employ some other minimization technique to determine the model coefficients. We calculate statics such as p-values and perform analyses of variance (ANOVA) to help us determine what is significant and what is not. Finally, promising parameters are subjected to tests of logic to make sure there is a reasonable explanation for why a certain parameter would turn out to be a significant "cost driver."

The underlying assumption behind the process described above is that if we have reasonably good data, use statistics appropriately, and have sound logic we will get a good cost model. But is that necessarily true? Regina Nuzzo, in a *Nature* article published online titled "How Scientists Fool Themselves – and How They Can Stop," makes the following assertion:

> In today's environment, our talent for jumping to conclusions makes it all too easy to find false patterns in randomness, to ignore alternative explanations for a result or to accept 'reasonable' outcomes without question – that is, to ceaselessly lead ourselves astray without realizing it.

In his book "The Signal and the Noise" Nate Silver quotes Tomaso Poggio, an MIT neuroscientist who studies how our brains process information. Here is what Poggio says: "The problem is that these evolutionary instincts sometimes lead us to see patterns when there are none." Thus, and this is not surprise to anyone with children, we come programmed from the womb to try and make sense of the world. This programming was very useful to our ancestors when, as they hunted for food on the African Savanna, had to make life or death decisions in circumstances were information was incomplete and experience was everything. Unfortunately, this programming can lead us astray when it comes to making sense of the modern world.

The need to make sense of our world combined with other biases in our thinking, as described in my paper "The Psychology of Cost Estimating," can lead us to make serious and consequential mistakes when developing our models. For example, the optimism bias can cause us to be overly confident that the patterns we see in the data are truly there. The confirmation bias can cause us to reject information that indicates a specific parameter may not be as logical as it appears. The bandwagon bias can lead to group think where we mutually support each other in reaching the same conclusion. These biases cause us to make mistakes that lead to models that might look good, but may not be good for us. In the next section I talk about some of the more common mistakes we make.

Sins of Cost Modeling

Our space systems data is messy. Despite our best efforts to collect and normalize NASA space flight hardware data in a consistent manner, the end result are data sets that are non-homogeneous. For

example, Exhibit 1 shows a graph of robotic spacecraft electrical power subsystem cost versus Beginning of Life (BOL) power. The data for these spacecraft were collected using a standard NASA data collection process called a Cost Analysis Data Requirements (CADRe) document. The data was normalized by the same team of cost engineers to adjust for inflation, program content, and differences in cost accounting. These missions are all in a low earth or a near earth orbit (lunar, L1, etc.). All these spacecraft use solar arrays for electrical power generation.

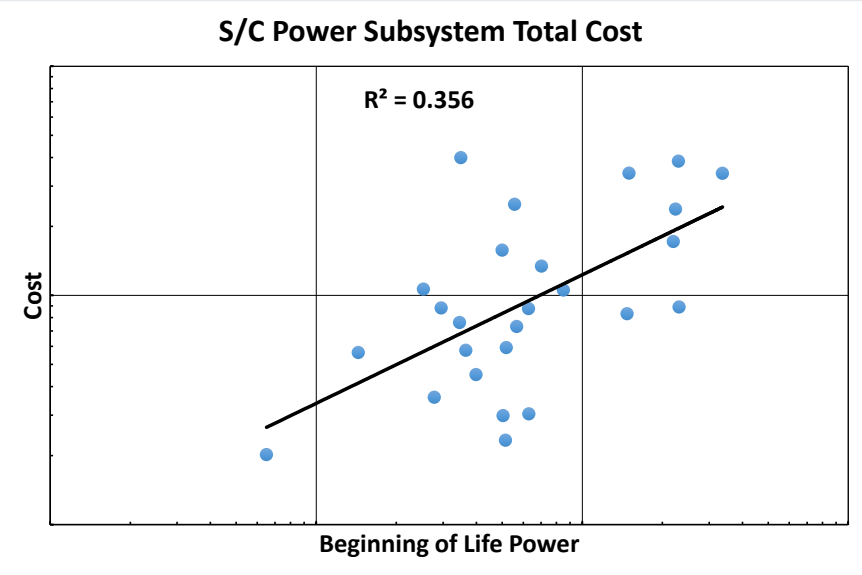

Exhibit 1. Spacecraft Electrical Power Subsystem Cost.

Despite the efforts to create a homogeneous data set, the amount of scatter in the data is large, in some cases more than an order of magnitude difference in cost for the same BOL power. The rather poor R^2 indicates a surprising lack of predictive power for such a logical technical parameter. Interestingly, the degree of scatter shown in Exhibit 1 is not unusual. In fact, it is quite normal for space systems cost data. Obviously, there are other factors at work here that have a significant influence on the cost.

The non-homogeneity in space systems cost data is driven by three factors. Number one, each NASA mission is a unique undertaking, with specific science requirements that often necessitate custom technical solutions. The second factor is the relatively small size and specialization of the space flight hardware industrial base. There are very few builders of robotic spacecraft and these builders rely on a small number of specialized companies to provide key components such as star trackers and solar array cells. Components and systems are made to order with considerable touch labor. Finally, space system data is data of opportunity. Unlike data collected in a laboratory or on a factory floor, cost modelers cannot manage the environment that creates the data or perform repeatable experiments under controlled conditions.

The lack of orderly behavior in space systems cost data creates a fertile environment for analysts to make well-intentioned mistakes. After all, we are only following our biological mandate to make sense of confusion, order out of disorder. But our attempts to solve the problem may actually cause us to do more harm than good. So while not specifically a sin itself, data non-homogeneity may be the root of all cost modeling evil.

The first sin I want to talk about is the sin of confusing correlation with causation. Most of the independent variables used in parametric models are associative or scaling, not causative. For example, weight is a common input to parametric models, yet every experienced analyst knows that in some cases reducing weight can actually increase cost (by going to more exotic materials or miniaturizing the electronics, for example). Most analysts are good at rejecting independent variables that have little or no relationship logical relationship to the cost. The danger comes in how we ascribe predictive power to those parameters that do past the logic test.

In my opinion, one of the worst terms in parametric cost modeling is "cost driver." Because our parameters are most often associate or scaling, they no more "drive" cost than your dog can drive your car. Unless the input parameter has a direct relationship to how the system is designed or built, we should banish the term "cost driver" from our lexicon. Otherwise, we are bestowing upon a parameter an undeserved and misleading title, a title that implies that one can control the predicted cost by controlling the parameter value.

Sometimes a model builder does not take into account the correlation between input parameters. Failure to consider this correlation can lead to inputs that are largely duplicative in their predictive power, add nothing to the predictive capability of the model, and in some case cause parametric coefficients to become unstable. When this happens multicollinearity is said to exist. A good test to determine if your model has multicollinearity is the Variance Inflation Factor (VIF). Generally, multicollinearity may be a problem if the VIF exceeds 4 or 5, and under no circumstances should it exceed 10. There are cases where stakeholders or others with a vested interest in a model insist that a certain parameter must be in the CER because they know (based on knowledge, experience, legend, etc.) that that parameter is what drives cost. When such parameters are included it can mislead customers into believing that certain decisions determine cost, when the reality is just the opposite.

Another way to bias your model is to focus on the aforementioned R^2 statistic. R^2 is a quick way to determine the goodness of your model, and every model builder wants to be able to show off a high R^2 value (like 0.9, or better yet, 0.95). R^2 is useful, but R^2 can be manipulated. There are two very easy ways to improve R^2. The first is to cherry pick the data by removing outliers. In some cases, there may be a legitimate reason for removing a data point, such as known issues with content or completeness of the cost data. In other cases there may not be a known problem but the data point is so far from the general trend that some unidentified issue must be causing it to be out of family. It is certainly appropriate to scrutinize outliers to make sure there is not a problem, but; it is also not appropriate to eliminate a data point simply because doing so increases the R^2 value. Something to consider if you find a problem with an outlier: you should apply the same level of scrutiny to your better behaved data points to make sure they don't have the same problem. Because we are using data of opportunity we must be careful to make sure that we are not seeing problem as unique when in fact it is systematic.

The second way to boost the R^2 value is to increase the number of independent variables. As the number of input variables approach the number of data points, the model appears to be explaining more and more of the variation in the data. This happens because each time you add a parameter, you are forcing the model to account some of the variation in the dependent variable. However, the reality is that you are probably explaining random noise in the data rather creating a model that better predicts cost.

The general term for this problem is overfitting. The challenge we in the cost profession face is that because our data sets are small and the data is noisy, we are prone to use overfitting as a way to achieve an acceptable R^2 value. But this behavior comes with consequences. In his book "The Signal and the Noise" Nate Silver makes the following point:

> Overfitting represents a double whammy: it makes our model look better on paper but perform worse in the real world. Because of the latter trait, an overfit model eventually will get its comeuppance if and when it is used to make real predictions. Because of the former, it may look superficially more impressive until then, claiming to make very accurate and newsworthy predictions and to represent an advance over previously applied techniques…But if the model is fitting noise, it has the potential to hurt the science.

Thus by overfitting we can easily make cost models that look good but give poor predictions, with the cost estimating and analysis profession being the ultimate loser. Fortunately a number of statistical tests (such as t-tests, p-values, sequential F-tests, etc.) exist so that it should be easy for the analyst to identify, as additional parameters are added to the model, when the point of diminishing returns has been reached.

The final sin I want to discuss is the misuse of subjective parameters. Subjective parameters seem to be especially favored in the cost estimating field because they are a handy way to model the otherwise unexplained randomness in the data. These parameters go by names such as new design, heritage, or complexity. They represent well-meaning attempts to improve the predictive power of a model, but in the end can mislead us and our customers.

The danger with subjective parameters comes from the judgment required to define the parameter and the judgment used in assigning subjective values (or ratings) to the data. Most subjective parameters have rather vague and amorphous definitions. Take "new design" for example. In the space business, almost everything flown in space for the last 40 years owes its design, at least in part, to something that has flown before. Yes, there are specific instances where something new was flown such as the first time gallium-arsenide solar cells were used or the first flight of the solid state data recorder. However these applications of new technology tend to be a rather small percentage of a total space system. In fact, our data shows that most spacecraft use well understood technologies packaged in new and unique ways. Therefore, determining the amount of "new design" in a space system becomes a highly subjective exercise with the outcome subject to biased thinking.

We misuse subjective parameters when we put them into our regression analyses and treat them the same as more objective system parameters. This creates the illusion that a judgment based parameter is equal in value to an objective parameter. This illusion is further reinforced when these parameters turn out to have a predictive value, as measured by the statistics, which is at least as good as the objective parameters. However, what we are really doing is cleverly modeling the noise. I will demonstrate how easy it is to do this in the following section.

To further our abuse of subjective parameters, once we incorporate them into our models we treat the output of those models deterministically. In other words, we act as if the reduction in model error through the incorporation of the subjective parameter is real, and not a as a new source of estimating error. One could argue that by doing a cost risk analysis we can address the uncertainty associated with

the subjective parameter value and get results similar to the model uncertainty without the subjective parameter. I, however; contend that the optimism bias keeps us from putting a sufficiently large range on the subjective parameter, thus creating a false sense of comfort in our estimate.

For example, in Exhibit 2 I have taken a chart from Christian Smart's paper "Covered with Oil: Incorporating Realism in Cost Risk Analysis." The s-curves in Exhibit 2 are based on actual risk analyses over time for a real NASA project, and demonstrate unequivocally how optimistic we can be, especially in the early phases of a project.

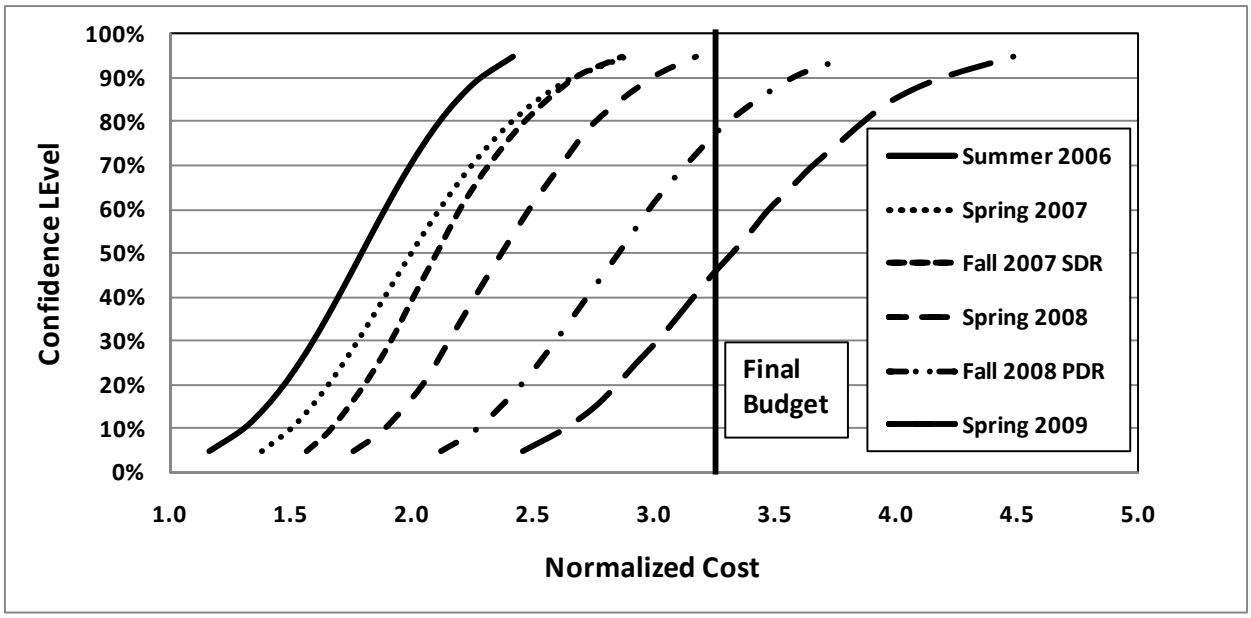

Exhibit 2. S-curves over time (Smart, 2015).

Not only does the s-curve move to the right by almost a factor of 2.5 in just a little more than two years, but the curve also begins to flatten, indicating that the analyst has less certainty about the outcome than in the earlier analyses. While this behavior is (fortunately) not typical of all NASA projects, analysis of NASA and DoD cost growth data (Smart 2010 and Prince 2015) shows that the average amount of cost growth on space and high tech projects is generally around 50%, significantly greater than the typical 30% or 35% reserve placed on most estimates.

The Lure of Subjective Parameters

Much like the "Dark Side of the Force" from "Star Wars" mythology, subjective parameters seduce the cost model developer. This seduction comes from their power to explain the random noise in our data, to improve the model statistics, and to enable the estimator to fine-tune the estimate to reflect their evaluation of a new system. To illustrate how this occurs look at the graph in Exhibit 3.

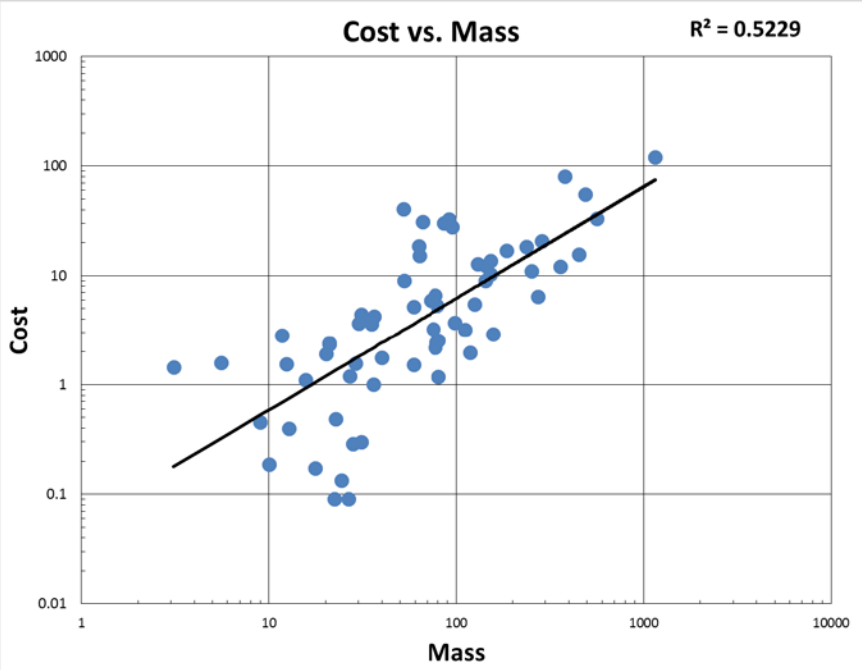

Exhibit 3. Cost versus Mass for a Spacecraft Subsystem.

Exhibit 3 shows the cost versus mass for a specific earth-orbiting spacecraft subsystem. Like the data shown in Exhibit 1, there is a significant amount of scatter (or noise). The R^2 value of 0.5229 shows that mass alone can only explain a little more than 50% of the variation in cost.

So let's see what happens when we use a subjective parameter to explain the noise in the data. In this case, the subjective parameter is called New Design. Our New Design scale has eight categories, with category 1 representing the data points with the greatest amount of new design, and category 8 representing the data points with the least amount of new design. The actual New Design values used for the categories range from 100% for category 1; followed by 91%, 79%, 64%, 43%, 25%, and 15% for categories 7 through 2, respectively, all the way down to 5% new design for category 8. Each of the data points is assigned to a New Design category is based on the analyst's judgment regarding the technology used, the design inheritance from previous spacecraft, and the overall state of the art at the time the subsystem was developed.

In Exhibit 4, a color coding scheme shows the result of assigning each data point to a New Design category.

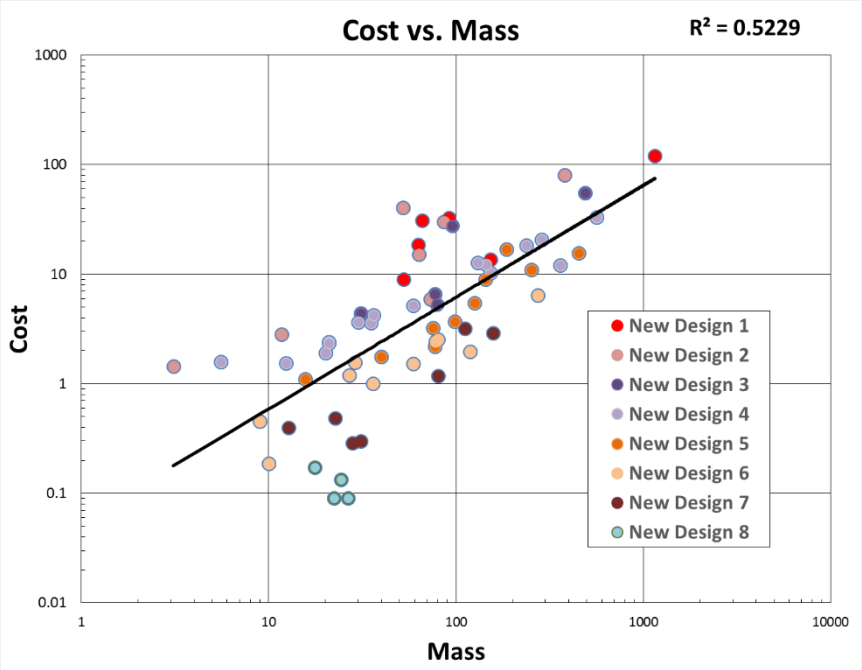

Exhibit 4: New Design Category Assignments for all Data.

In Exhibit 4 the trend is clear. While there is substantial overlap between categories, in general, each category represents a reduction in cost as compared to the trend line. And, with the exception of a total overlap between categories 1 and 2, the average cost per pound decreases consistently from category to category. To the casual observer, this New Design categorization seems logical and reasonable. In practice it proves to be a powerful tool for building a better cost model.

Look at Exhibit 5. Exhibit 5 shows the power of subjective parameters and why it can be so hard to avoid using them. With an R^2 value of 0.9245 we have gone from a mediocre mass-based CER to wonderful two parameter CER, without any cherry picking of the data or overfitting the model. In addition the statistics are excellent, as can be seen in Exhibit 6.

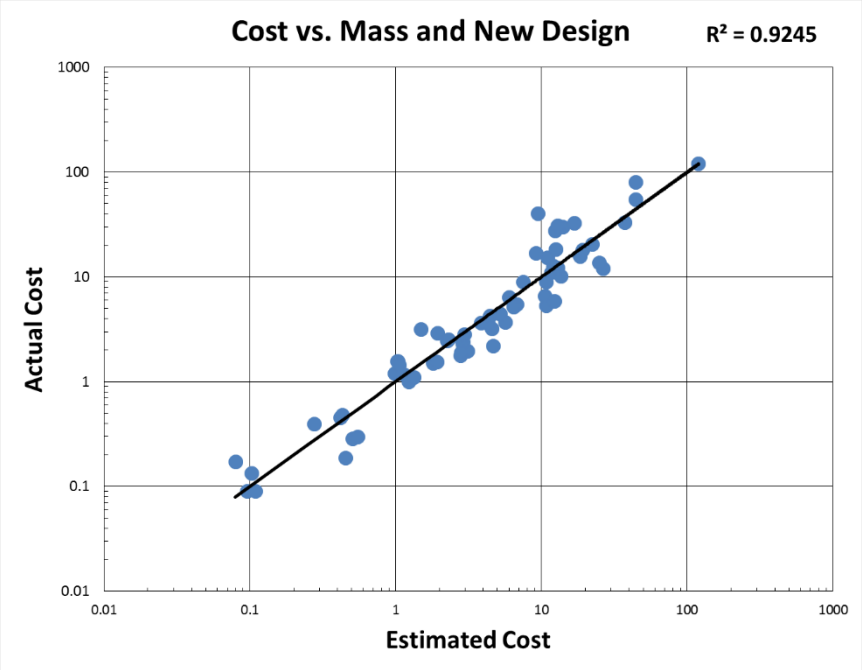

Exhibit 5: CER with Mass and New Design.

Regression Statistics	
Multiple R	0.9615
R Square	0.9245
Adjusted R Square	0.9220
Standard Error	0.4658
Observations	65

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	2	164.6983	82.3492	379.4792	0.0000
Residual	62	13.4544	0.2170		
Total	64	178.1527			

	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	-0.7020	0.2427	-2.8928	0.0053
ln(weight)	0.7783	0.0510	15.2471	0.0000
ln(newdesign)	1.3560	0.0747	18.1579	0.0000

Exhibit 6: Regression Statistics for Mass and New Design CER.

You can see why the lure of using subjective variables is so strong. Why live with noisy data and a mediocre CER when you can use a logical rating scheme to create a significantly improved model? Better yet, you get a model that produces expected results: the estimator can now show the cost savings from using an existing design or the cost penalty for incorporating a new technology. But when we go down this path we are giving in to our biases and giving in to bad modeling practices, practices that create models that deceive us and we can use to deceive our customers.

For those of you who continue to hold on to the belief that a good cost engineer can develop meaningful subjective parameters that explain the noise in our data, let me offer the following example. Exhibit 7 shows an actual versus estimated graph using the same data set as Exhibit 5, but different values for the new design parameter.

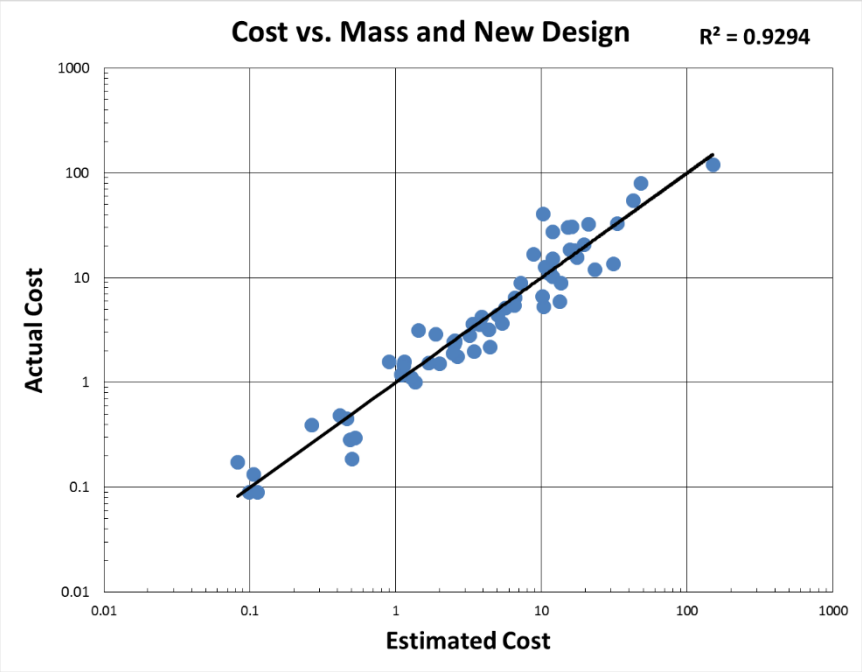

Exhibit 7. Cost versus Mass and New Design with Different New Design Parameter Values.

As you can see, the graphs in Exhibits 5 and 7 look very similar and the R^2 values for each are almost identical (detailed statistics are provided in Appendix A). From that visual examination one could conclude that since the cost and mass values are the same, the new design parameter values much be similar. Now look at the table in Exhibit 8.

New Design Category	Exhibit 5 Parameter Values	Exhibit 7 Parameter Values
Category 1	1.00	8
Category 2	0.91	7
Category 3	0.79	6
Category 4	0.64	5
Category 5	0.43	4
Category 6	0.25	3
Category 7	0.15	2
Category 8	0.05	1

Exhibit 8: New Design Category Parameter Values.

Exhibit 8 shows the parameter values for the eight New Design Categories. Look at the values in the column titled “Exhibit 5 Parameter Values.” These purport to represent the percentage of new design

present in the subsystem. Thus for a New Design Category 2, 91% (0.91) of the subsystem is a new design. For a category 3, 79% is a new design, and so on. Look at the values in the last column, the one titled "Exhibit 7 Parameter Values." These are the values used in the regression for Exhibit 7. Note that they are integers ranging from a value of 8 for the largest amount of new design to a value of 1 for the least. This is an ordinal scale, representing the order of the new design categories.

Now look again at the scatter of the data and the R^2 values in Exhibits 5 and 7. They are almost identical. Yet the values used for the New Design parameter are drastically different. How can this be? The reason both values work is that the power of this particular New Design parameter lies not in its assigned value, but in the fact that it provides a categorization of the data in a systematic and logical way that accounts for the noise. Look back at Exhibit 5. If you start with New Design Category 8 and work your way to New Design Category 1, you will see the cost of the subsystem increases with each category. There is significant overlap, but the trend is clear.

What we are seeing is a case of imposing an arbitrary structure on random noise. The New Design parameter categories and values are a sincere attempt to capture what we intuitively know about high technology systems: the greater the amount of new design the higher the cost, and vice versa. The model looks good and it behaves correctly, yet is has no more predictive power than a simple mass-based model, and can easily fool us and our customers.

The Practice of Self-Deception

There are several biases that affect how we develop our models. These biases come into play whether or not we use subjective parameters. However, subjective parameters make it easier for these biases to affect our decision making and our models.

Several of these biases I discussed in my paper "The Psychology of Cost Estimating." Others were identified by Regina Nuzzo in the article I quoted earlier. Below is a short list of biases that can contribute to deceptive modeling practices along with their definitions.

Asymmetric Attention: When we give expected outcomes little scrutiny yet rigorously check non-intuitive results we are displaying Asymmetric Attention. The problem caused by Asymmetric Attention is that is can lead us to overlook errors in our favor. For example, Asymmetric Attention is likely to cause us to pay limited attention to a subjective parameter like our New Design variable as long as that parameter behaves in a way we expect.

Representativeness: Representativeness is our tendency to relate something new or novel to something we know or are familiar with. The representative bias gets us into trouble by causing us to see patterns in randomness and to underappreciate how random a random outcome can truly be. Therefore our tendency is to try to impose an arbitrary structure on random noise. The fact that such a structure works not a justification for doing so.

What you see is all there is (WYSIATI): WYSIATI is a phrase coined by Daniel Kahneman to describe how our minds can quickly develop a coherent story out of limited information. Two surprising facts emerge from WYSIATI. First, the less information we have the more confident we are in our coherent story, especially a story about a subjective parameter. Second, the coherent stories that we build often ignore probability and statistics.

Halo/Horns Effect: The Halo/Horns effect (also known as the confirmation bias) is our tendency to emphasize data that agrees with our belief or intuitive assessment, and to discount information that disagrees with our position. The Halo/Horns effect can also cause us to look for (or be more open to accepting) data that confirms our position or opinion. Obviously, the danger with this bias is that we will overlook or discount important information that is inconsistent with the desired outcome.

Plausibility Effect: When we believe the more plausible outcome over the more probable outcome, we are falling victim to the plausibility effect. The Plausibility Effect occurs because we like explanations that address all of the facts, even if those facts are the result of random events. In other words, we like stories that rationalize the results. Cost modelers fall victim to the Plausibility Effect whenever we confuse a good story about the data rather than letting the data speak for itself. A related bias, called *Storytelling*, speaks to how we will find stories to rationalize the results we see.

Attractiveness: Appearances matter. Psychologists have known for years that people assign more favorable characteristics to attractive people or products. We are also more likely to believe a good presenter over a poor presenter. Attractiveness and the Plausibility Effect and the Confirmation Bias are interrelated. We like a good (attractive) story that makes sense and explains all the variation in the data, especially if it confirms a previously held belief or opinion.

A Bias-Free Cost Model

Now that understand the problem of biases and subjective parameters, let's build a model using only objective parameters. For this particular model we have a small data set, only nine points. Also, there is significant scatter in the data, as can be seen in Exhibit 9.

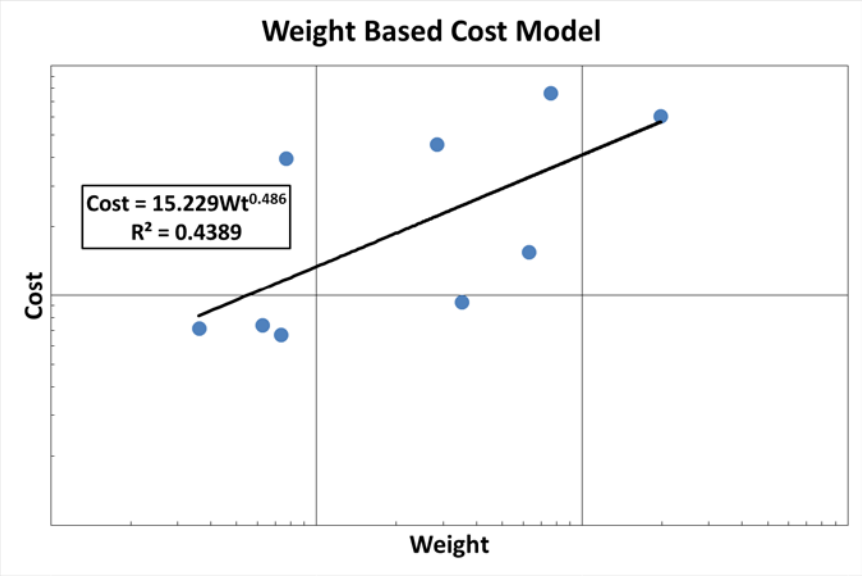

Exhibit 9: Simple Weight-Based Cost Model.

While the trend is reasonable, the significant scatter in the data results in a rather poor R^2 of 0.4389. The danger at this point, as we have learned, is that we could use a bad modelling practice such as overfitting the data or using a subjective parameter. But what if we found a second objective parameter

that provided substantial predictive power? Exhibit 10 shows the impact of adding a date parameter to the model, in this case an indicator variable identifying data points developed before and after 1970.

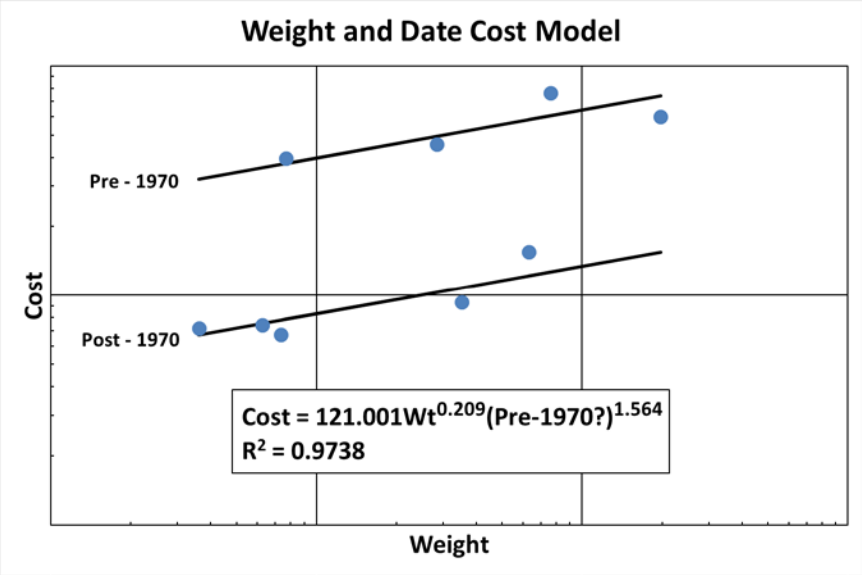

Exhibit 10: Weight and Date Cost Model.

Adding the indicator variable significantly improves the quality of the CER (a more complete listing of the pertinent statistics for both CERs is given in Appendix B). But the question needs to be asked: why 1970? Space systems developed prior to 1970 were pushing the art in terms of manufacturing capability, technology, and were often schedule driven. After 1970 advancing computer technology enabled better design processes, with fewer test articles and systems tests. In general the technology was better understood, schedule pressures were reduced, and the many failures of the early Space Age created a large knowledge base of what did and did not work. Thus by both statistics and logic, the CER shown in Exhibit 10 is far superior to the CER in Exhibit 9.

Exhibit 11 compares some early estimates for a new space system similar to those in our model. If we are using our model to validate these early estimates, which are a mixture of parametric and engineering build-up, we would probably say they are conservative, since they are all above the post-1970 trend line.

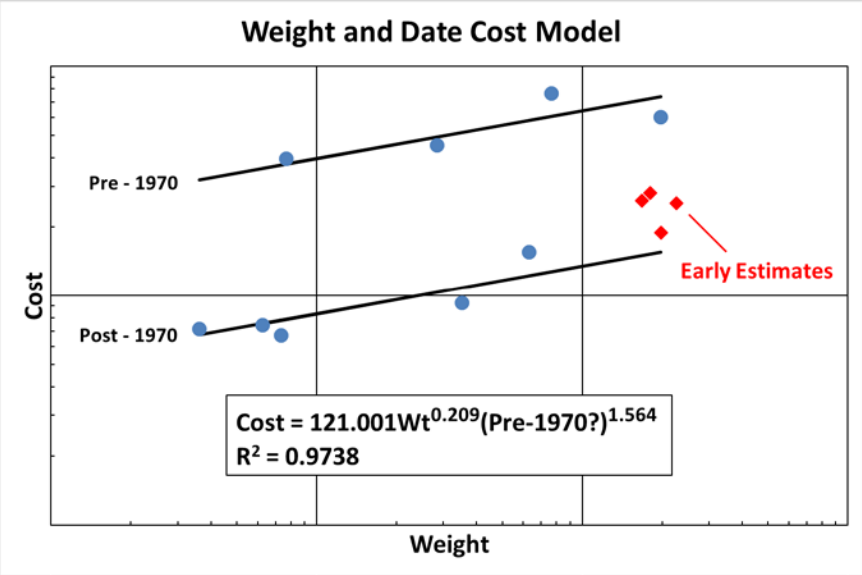

Exhibit 11: Comparison of Model to Early Estimates.

Exhibit 12 shows the same plot with more recent cost estimates. These more recent estimates are approaching the pre-1970 trend line. Obviously, the project did not get the message that anything developed after 1970 should not cost this much. So what went wrong, why did this model, which was built on objective information, which is not over-subscribed, which has no biases, fail?

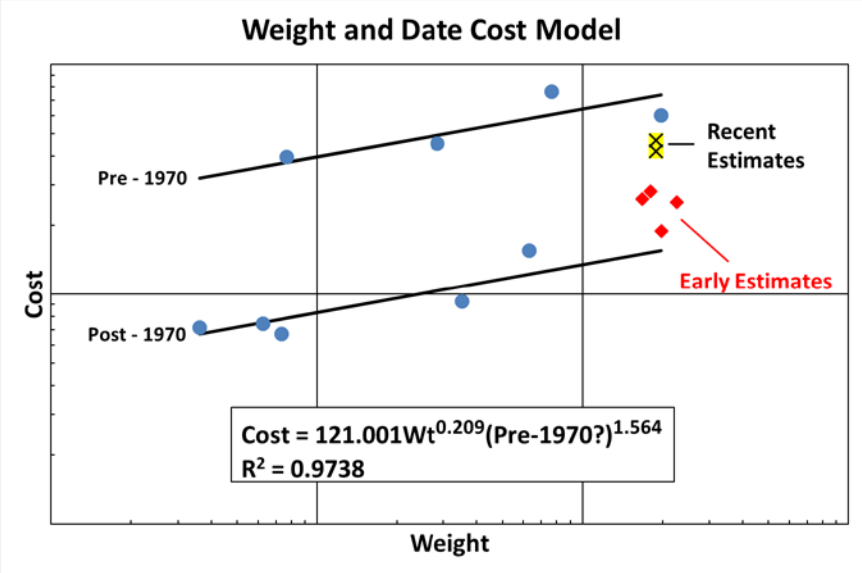

Exhibit 12: Comparison of Model to Recent Estimates.

Take a look at Exhibit 13. In Exhibit 13 I have numbered the historical data points and identified one or more pertinent characteristics about each one. Note that six of the nine data points are first of a kind systems but that the three lowest cost data points (5, 6, & 7) contain significant heritage from data point 1. Data points 5, 6, 7, and 8 were developed to interface with a pre-existing system, simplifying development. Data point 9 is significantly less complex than the other eight. In terms of size and

function, data points 3 and 4 are the closest analogs to the new system we are developing. Therefore, we should not be surprised that the cost appears to be gravitating towards these analogs.

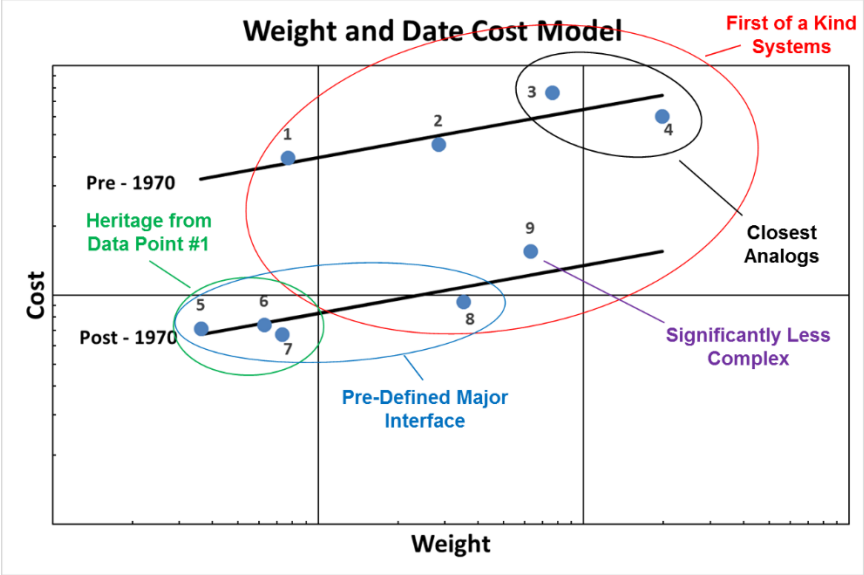

Exhibit 13: A Different View of the Model Data.

What is striking about Exhibit 13 is that is illustrates a story about the data that is much more complex and nuanced than the two variable model. In other words, each data point has a story to tell. And when we develop parametric models we either ignore these stories because it is difficult to synthesize them into a single number or we oversimplify them to create a representative value for modeling.

So what is a responsible cost estimator supposed to do? If she tries to address the uniqueness of the data, she risks over-specifying the model or creating unsupportable subjective parameters. If she sticks with simple objective parameters, she may still fall victim to biases, and the model may yield unsatisfactory results. The solution to this problem lies not in our approach to modeling, but in our approach to how we use our models, and how we use our data.

Parametric Cost Estimating and Data

Parametric cost estimators are defined by their models, but in reality it is our data that separates us from other approaches to cost estimating (one could argue, quite successfully I believe, that all cost estimating is data driven, but that is a topic for another paper). A model is simply a representation of reality, and in our particular situation, given the amount of random noise in our data, a rather pale representation at that. Therefore, to be an accomplished parametric cost estimator, you must know and understand the historical data that is the basis for your model(s) and that is most relevant to the system you are costing. You cannot separate parametric cost estimating from a knowledge and understanding of the data!

Parametric cost models are valuable and useful tools, but they must be used intelligently. The power of a parametric cost model comes not from its statistical prowess, but from its ability to enable to analyst to extrapolate from the known to the unknown. This is why the connection between the data and the model is so important. Using a model free of the data can give you a reasonable answer, but if the

model is the only basis for your estimate then you may be on thin ice. Relevant data should be the starting point for your cost estimate and it should also form the foundation for validating your estimate. A good parametric model used in conjunction with relevant data makes for an estimate that is credible, supportable, and defendable.

Making a Better Cost Model

When you are working with data of opportunity, creating a parametric cost model is hard. NASA data is full of random noise, noise that comes from how the data was collected and analyzed, but more importantly, noise from all of the random events that happen during a system's development. Often the parametric cost modeler is faced with a choice: engage in R^2 boosting practices such as cherry picking the data or overfitting the model; use subjective parameters to explain the noise; or embrace the messiness, the noise, the randomness. My recommendation is that you embrace the mess.

Embracing the mess means that you accept that there are limitations as to how far our models can go in explaining the underlying data as well as the cost of future systems. It also means that you are going to rely on the story behind the data to explain the variation in the data and support your cost estimate. You are going to honor what the data is telling you, not try to ignore it or oversimplify it or explain it away. When you embrace the mess you will let the data guide your model and your estimate. The data will tell you if your cost should be higher or lower. The data will provide the foundation.

To make a better model you must learn to question everything. We are good at questioning non-intuitive results, but we must learn to question our intuitive results as well. A correct intuitive result may be right for the wrong reason. An excellent example is the New Design parameter we investigated earlier. The percent new design values of 5%, 15%, 25%, etc. gave intuitive results when used in the regression analysis. But as I showed in Exhibit 7, it was not the new design values but rather the stratification of the data (and the imposition of an arbitrary structure on the randomness) that led to these results.

The flip side of not questioning intuitive results is the attempt to make non-intuitive results intuitive. Here is how it works. You are developing a new hardware CER. The technical people tell you that parameter X is the most important parameter for determining the cost. You regress cost against parameter X and you get lousy results. The typical human response is to try and find some causative factor (bad data, outliers, etc.) to explain the non-intuitive outcome. But it could also be true that the technical experts are doing their own storytelling, that they have developed their own mythology based on representativeness to explain what drives cost. It is also possible that parameter X is an important input into how the technical expert does their job, but not to the overall cost of the system (Kahneman's WYSIATI).

Another idea for making better models is to have the process and the results reviewed by an independent, non-advocate team. Getting an independent review can raise questions you did not think about and can force you to defend your choices. In the process of explaining your rationale to an independent party problems in logic or approach may be discovered. An independent reviewer may also suggest an alternative approach, explanation, or technique that you did not consider. And don't limit your review to the process and results, have someone check the math. Mathematical mistakes that don't create non-intuitive results may not be caught, invalidating an outcome.

A final suggestion for improving our models is to have a different team (or analyst) take the data and develop their own model. If the independent analysts reach similar conclusions regarding input variables, transformations, and minimization techniques; then that can validate your model. Even if the independent team reaches different conclusions that does not necessary invalidate your work, but it will require that you re-examine your process and approach to ensure that you have not biased the outcome, overlooked an alternative explanation for the behavior of the data, or made a mistake.

I realize that resources (both time and people) may make the independent development team approach difficult, if not impossible, to implement. However, at NASA Marshall Space Flight Center (MSFC) we are making the data we use to develop our CERs for the Project Cost Estimating Capability (PCEC) available to the entire NASA cost community. We are hoping that other cost analysts will take the data and perform their own model developments. I realize that within the larger parametric cost community there are different schools of thought concerning equation forms, minimization criteria, etc. If two different groups approach their analysis of a data set from very different mindsets, it could lead to an interesting comparative analysis where we all learn something unexpected, and possibly useful.

Improving Model Accuracy

We have become quite adept at creating parametric models that reproduce the underlying data with a reasonable level of accuracy. But is that a real measure of the quality of the model? After all, the purpose of our models is to help us estimate the cost for something that has never been built. What if we measured the quality of our cost models by how well they estimate the cost for these new systems?

In the space business development cycles are often long, on the order of 4 to 8 or more years. The number of system changes made between the initial estimates and what it actually built also make it difficult to get useful feedback on how well a model performs. Still, the best way to determine model performance is to use it to make a prediction and then see how well that prediction turns out. Benchmarking past estimates to actuals should be standard operating procedure for all cost organizations, both as a way to evaluate the model(s) and (perhaps more importantly) as a way to evaluate the organization's estimating process.

Another approach to improving model accuracy is to use one or more resampling techniques, such as bootstrapping, cross-validating, or jackknifing. Bootstrap resampling involves drawing a random sample from your data set, with replacement, that is the same size as your model data set. In other words, if your model is based on N data points, you will draw a sample of size N from your model data, replacing (or returning) the drawn data point each time. By using Monte Carlo simulation to run thousands of trials, you can use bootstrapping to estimating the confidence interval on your model coefficients as well as a prediction interval for your estimates.

Jackknifing involves recalculating the model coefficients N-1 times (N being the size of you data set), each time leaving out a different data point. Using the Jackknife technique you can calculate a mean and standard deviation for your model coefficients as well as the amount of bias in your baseline model. The Jackknife technique is sometimes preferred over bootstrapping due to its ease of calculation and repeatable results.

Cross-validation is a technique particularly well suited for assessing how well a cost model will perform in the real world. In cross-validation you divide your data set into a training (or known) data set and a

validation (or unknown) data set. Typically this is done multiple times, each time using different data points for the training and validation data sets. There are various methods for selecting the size and composition of the training and validation sets (Wikipedia lists several). Since the purpose of cross-validation is to measure model performance, you can calculate statistics such as the mean squared error or the mean absolute deviation to determine estimating error.

A third approach to improving model accuracy is to benchmark a model against new data. Obviously, the new data should be of the same type used to develop your model (i.e. using a spacecraft data to benchmark a spacecraft cost model). The technique is simple. Take the cost for a recently completed system, normalize the cost as necessary so that it is consistent with your model output, develop an estimate for the system using your model, and compare the results to the actuals. The advantage of this approach is that you get unbiased feedback on how your model performed. If the costs are reasonably close (I will leave it up to you to define "reasonable") then your model is doing good, if the costs are not close, then your model may be lacking in some way.

Two things to consider when benchmarking against new data. While a poor result indicates your model accuracy may be low, a good result is not proof that your model is accurate. In other words, good results do not guarantee continued good results. The other problem with benchmarking is that unless your model estimates something that produces large numbers of actuals, you will not have sufficient data to develop a statistical estimation of model performance. This is a real problem at NASA, where even in a good year we may only launch 5 or 6 science missions and decades can elapse between the development of human spaceflight systems.

Final Thoughts and Conclusions

Building a parametric cost model is hard work. The data is noisy and often does not behave like we want it to. We need statistics to give us an indication of the goodness of our models, but; statistics can be manipulated and mislead. On top of all of that, our own very human biases can lead us astray; causing us to see patterns in the noise and draw false conclusions from the data.

Yet, it is the data itself that is the foundation for making better cost estimates and cost models. I believe the mistake we often make is we believe that our models are representative of the data; that our models summarize the experiences, the knowledge, and the stories contained in the data. However, it is the opposite that is true. Our models are but imitations of reality. They give us trends, but not truth. The experiences, the knowledge, and the stories that we need in order to make good cost estimates is bound up in the data. You cannot separate good cost estimating from a knowledge of the historical data.

One final thought. It is our attempts to make sense out of the randomness that leads us astray. In order to make progress as cost modelers and cost estimators, we must accept that there are real limitations on our ability to model the past and predict the future. I do not believe we should throw up our hands and say this is the best we can do. Rather, to see real improvement we must first recognize these limitations, avoid the easy but misleading solutions, and seek to find ways to better model the world we live in. I don't have any simple solutions. Perhaps the answers lie in better data or in a totally different approach to simulating how the world works. All I know is that we must do our best to speak truth to ourselves and our customers. Misleading ourselves and our customers will, in the end, result in an inability to have a positive impact on those we serve.

Bibliography

Ariely, Dan, *Predictably Irrational*, Revised and Expanded Edition, New York: Harper Perennial, 2009

Aschwanden, Christie, "Your Brain is Primed to Reach False Conclusions." *fivethirtyeight.* February 17, 2015. <http://fivethirtyeight.com/features/your-brain-is-primed-to-reach-false-conclusions/

Dickson, Paul, *The Official Rules,* New York: Dell Publishing, 1978

Gladwell, Malcolm, *Blink, The Power of Thinking Without Thinking*, New York: Little, Brown and Company, 2005

Hamaker, Joseph W., "What Are Quality Cost Estimates? Or the 260 Hz Cost Estimate," *Journal of Parametrics* Vol. 25, Issue No. 1, 2007: 1 – 7

Hubbard, Douglas W., *How to Measure Anything*, New Jersey: John Wiley & Sons, 2010

Kahneman, Daniel, *Thinking, Fast and Slow*, New York: Farrar, Straus and Giroux, 2011

Levitt, Steven D. and Dubner, Stephen J., *Freakonomics, a Rouge Economist Explores the Hidden Side of Everything*, New York: Harper Perennial, 2009

Mlodinow, Leonard, *The Drunkards Walk: How Randomness Rules Our Lives*, New York: Pantheon Books, 2008

Mooney, Chris, "The Science of Why We Don't Believe Science." *Mother Jones.* May/June 2011. <http://www.motherjones.com/politics/2011/03/denial-science-chris-mooney>

Nuzzo, Regina, "How scientists fool themselves – and how they can stop." Nature. October 7, 2015. <http://www.nature.com/news/how-scientists-fool-themselves-and-how-they-can-stop/>

Prince, Frank, "The Psychology of Cost Estimating," *Proceedings of the 2015 International Cost Estimating and Analysis Association Professional Development and Training Workshop*, San Diego, June, 2015

Siegel, Eric, "The One Rule Every Data Scientist (and Manager) Should Know By Heart," GovExec.com, December 21, 2015. <http://www.govexec.com/technology/2015/12/oneruleeverydatascientistandmanagershouldknowheart/124803/print/>

Silver, Nate, *The Signal and the Noise: Why most Predictions Fail but some Don't*, New York: The Penguin Press, 2012

Smart, Christian, "Bayesian Parametrics: How to Develop a CER with Limited Data and Even Without Data," *Proceedings of the 2014 International Cost Estimating and Analysis Association Professional Development and Training Workshop*, Colorado: June, 2014

Smart, Christian, "Covered in Oil, Realism in Cost Risk Analysis," *Journal of Cost Analysis and Parametrics,* Vol. 8, Issue No. 3, 2015: 186-205.

Surowiecki, James, *The Wisdom of Crowds*, New York: Anchor Books, 2005

Thayer, Richard H. and Sunstein, Cass R., *Nudge: Improving Decisions About Health, Wealth, and Happiness,* New York: Penguin Books, 2009.

Appendix A

Detailed Statistics for Exhibit 5.

Regression Statistics	
Multiple R	0.9615
R Square	0.9245
Adjusted R Square	0.9220
Standard Error	0.4658
Observations	65

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	2	164.6983	82.3492	379.4792	0.0000
Residual	62	13.4544	0.2170		
Total	64	178.1527			

	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	-0.7020	0.2427	-2.8928	0.0053
ln(weight)	0.7783	0.0510	15.2471	0.0000
ln(newdesign)	1.3560	0.0747	18.1579	0.0000

Appendix B

Detailed Statistics for Exhibit 7.

Regression Statistics	
Multiple R	0.9640
R Square	0.9294
Adjusted R Square	0.9271
Standard Error	0.4505
Observations	65

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	2	165.5706	82.7853	407.9359	0.0000
Residual	62	12.5821	0.2029		
Total	64	178.1527			

	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	-4.7301	0.2263	-20.9025	0.0000
ln(weight)	0.7784	0.0493	15.7774	0.0000
ln(newdesign)	2.0452	0.1083	18.8909	0.0000